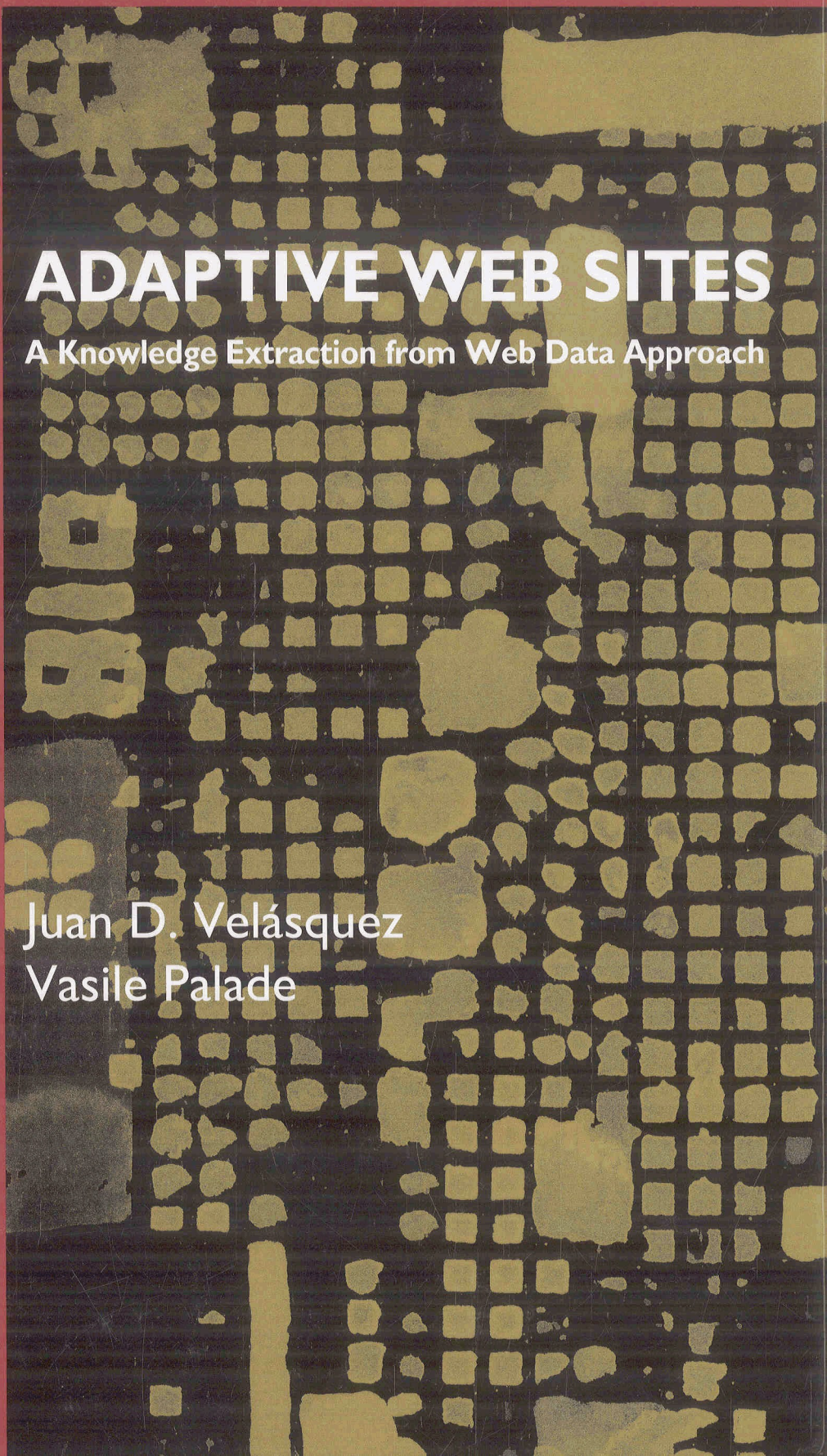
Frontiers in Artificial Intelligence and Applications



ADAPTIVE WEB SITES

A Knowledge Extraction from Web Data Approach

Juan D. Velásquez
Vasile Palade

IOS Press

Contents

1 Introduction **1**

1.1 The World Wide Web 2

1.2 Towards new portal generation 5

1.3 Structure of the book 7

2 Web data **9**

2.1 Web's Operation 10

2.2 The information behind the clicks 13

2.2.1 Session reconstruction process 15

2.2.2 Finding real sessions 18

2.3 The information contained in a web page 19

2.3.1 Web page content 19

2.3.2 Web page links 21

2.4 Summary 23

3 Knowledge discovery from web data **25**

3.1 Overview 26

3.2 Data sources and cleaning 28

3.3 Data consolidation and information repositories 30

3.4 Data Mining 32

3.4.1 Motivation 32

3.4.2 Data Mining techniques 33

3.4.2.1 Association rules . . . 33
3.4.2.2 Classification . . . 34
3.4.2.3 Clustering . . . 34
3.5 Tools for mining data . . . 37
3.5.1 Artificial Neural Networks (ANN) . . . 37
3.5.2 Self-Organizing Feature Maps (SOFMs) . . . 41
3.5.3 K-means . . . 43
3.5.4 Decisions trees . . . 44
3.5.5 Bayesian networks . . . 46
3.5.6 K-Nearest Neighbor (KNN) . . . 48
3.5.7 Support vector machines (SVMs) . . . 50
3.6 Using data mining to extract knowledge . . . 53
3.7 Validation of the extracted knowledge . . . 55
3.8 Mining the web . . . 55
3.9 Summary . . . 56

4 Web information repository **59**

4.1 A short history of data storage . . . 60
4.2 Storing historical data . . . 62
4.3 Information systems . . . 63
4.4 Data Mart and Data Warehouse . . . 65
4.4.1 The multidimensional analysis . . . 67
4.4.2 The Cube Model . . . 70
4.4.3 The Star Model . . . 72
4.4.4 The Extraction, Transformation and Loading Process . . . 75
4.4.4.1 Extraction . . . 75
4.4.4.2 Transformation . . . 76
4.4.4.3 Loading . . . 77
4.5 Web warehousing . . . 77

4.6 Information repository for web data 79
4.6.1 Thinking the web data in several dimensions 80
4.6.2 Hyper cube model for storing web data 82
4.6.3 Star model for storing web data 84
4.6.4 Selecting a model for maintaining web data 85
4.6.5 ETL process applied to web data 86
4.6.5.1 Processing web page text content 87
4.6.5.2 Processing the inner web site hyperlinks structure . . . 87
4.6.5.3 Processing the web logs 88
4.7 Summary 90

5 Mining the Web **93**

5.1 Mining the structure 94
5.1.1 The HITS algorithm 95
5.1.2 The Page Rank algorithm 98
5.1.3 Identifying web communities 101
5.2 Mining the content 102
5.2.1 Classification of web page text content 103
5.2.2 Clustering for groups having similar web page text content . . . 105
5.2.3 Some applications 106
5.2.3.1 WEBSOM 106
5.2.3.2 Automatic web page text summarization 107
5.2.3.3 Extraction of key-text components from web pages . . 108
5.3 Mining the usage data 109
5.3.1 Statistical methods 110
5.3.2 Clustering the user sessions 110
5.3.3 Classification of the user behavior in a web site 112
5.3.4 Using association rules for discovering navigation patterns . . . 113

5.3.5 Using sequence patterns for discovering common access paths . 114
5.3.6 Some particular implementations 115
5.3.6.1 Web query mining 115
5.3.6.2 Prefetching and caching 116
5.3.6.3 Helping the user's navigation in a web site 118
5.3.6.4 Improving the web site structure and content 119
5.3.6.5 Web-based adaptive systems 120
5.4 Summary 121

6 Web-based personalization systems **125**

6.1 Recommendation Systems 126
6.1.1 Short historical review 127
6.1.2 Web-based recommender systems 129
6.1.2.1 Web recommender systems, particular approaches and examples 132
6.2 Systems for personalization 133
6.2.1 Computerized personalization 134
6.2.2 Effectiveness of computerized personalization systems 136
6.2.3 Computerized personalization approaches 137
6.3 Web personalization 139
6.3.1 Aspects of web personalization privacy 141
6.3.2 Main approaches for web personalization 143
6.3.3 Privacy aspects of web personalization privacy 144
6.4 Adaptive web-based systems 147
6.4.1 A short introduction 148
6.4.2 Elements to take into account 150
6.4.3 Web site changes and recommendations 151
6.4.4 Adaptive systems for web sites 153
6.5 Summary 154

7 Extracting patterns from user behavior in a web site 157

7.1 Modelling the web user behavior . . . 158

7.2 Web data preparation process . . . 162

7.2.1 Comparing web page contents . . . 163

7.2.2 Comparing the user navigation sequences . . . 166

7.3 Extracting user browsing preferences . . . 169

7.3.1 Comparing user browsing behavior . . . 169

7.3.2 Applying a clustering algorithm for extracting navigation patterns 170

7.4 Extracting user web page content preferences . . . 173

7.4.1 Comparing user text preferences . . . 174

7.4.2 Identifying web site keywords . . . 175

7.5 Summary . . . 177

8 Acquiring and maintaining knowledge extracted from web data 179

8.1 Knowledge Representation . . . 180

8.1.1 Fundamental roles of knowledge representation . . . 180

8.1.2 Rules . . . 182

8.1.3 Knowledge repository . . . 183

8.2 Representing and maintaining knowledge . . . 183

8.3 Knowledge web users . . . 185

8.4 A framework to maintain knowledge extracted from web data . . . 186

8.4.1 Overview . . . 186

8.4.2 The Web Information Repository . . . 188

8.4.3 The Knowledge Base . . . 189

8.4.3.1 Pattern Repository . . . 190

8.4.3.2 Rule Repository . . . 191

8.5 Integration with adaptive web sites . . . 193

8.6 Summary . . . 194

9 A framework for developing adaptive web sites **195**

9.1 The adaptive web site proposal 196

9.2 Selecting web data 197

9.3 Extracting information from web data 200

9.3.1 The star model used for the creation of the WIR 201

9.3.2 Session reconstruction process 201

9.3.3 Web page content preprocessing 206

9.4 Applying web mining techniques 208

9.4.1 Analyzing the user browsing behavior 208

9.4.1.1 Applying statistics 208

9.4.1.2 Using SOFM for extracting navigation patterns 209

9.4.1.3 Using K-means for extracting navigation patterns . . . 212

9.4.2 Analyzing user text preferences 213

9.5 Using the extracted knowledge for creating recommendations 217

9.5.1 Offline recommendations 217

9.5.1.1 Structure recommendations 217

9.5.1.2 Content recommendations 219

9.5.2 Online recommendations 220

9.5.3 Testing the recommendation effectiveness 220

9.5.3.1 Testing offline structure recommendation 221

9.5.3.2 Testing offline content recommendation 225

9.5.3.3 Testing online navigation recommendation 226

9.5.4 Storing the extracted knowledge 229

9.5.4.1 Pattern Repository 230

9.5.4.2 Rules for navigation recommendations 231

9.6 Summary 233

In place of conclusions **237**